Facial Expression Detection In Video-Recorded Images Using a Mobilenet-Based Transfer Learning Approach

Sultan Adam Maulana*1
1STMIK AMIKBANDUNG, Bandung, Indonesia
E-mail: *1 **sulthonadammaulana@gmail.com**

Abstract

Emotions play an important role in human communication, and facial expressions are one of the main indicators for recognizing emotional states. Most studies in Facial Expression Recognition (FER) still focus on static images or real-time webcam tracking, while evaluation approaches based on recorded video remain less explored. This study aims to design a simple but functional pipeline to evaluate the performance of MobileNetV2 with transfer learning on verbal interaction video data. The Karolinska Directed Emotional Faces (KDEF) dataset was used for training with seven basic emotion classes, while the test data came from video recordings processed frame-by-frame. The pipeline includes frame extraction, face detection using Haar Cascade, image preprocessing, and classification with the fine-tuned MobileNetV2 model. Evaluation metrics such as accuracy, precision, recall, and F1-score were applied. The results show that the model reached 87% validation accuracy and was able to identify dominant emotions in video, although predictions tended to be biased toward the neutral class in subtle expressions such as anger and disgust. On the other hand, clearer expressions such as happy were detected more reliably. In conclusion, the proposed pipeline successfully bridges static-image models with video data, offering a practical and efficient evaluation approach that can support Human-Computer Interaction (HCI) applications on resource-limited devices.

Keywords: *Facial Expression Recognition, MobileNetV2, Transfer Learning, Verbal Interaction, Video Analysis*

1. INTRODUCTION

Human interaction relies not only on words, but also on facial expressions that can convey emotional states. Research in the field of *facial Expression Recognition* (FER) has evolved rapidly along with the advancement of deep learning, especially *Convolutional Neural Networks* (CNN). However, most research is still limited to static or *real-time images. Tracking with a webcam,* so that video recording-based evaluation is still rarely done [1][2].

MobileNetV2, as a lightweight CNN architecture with *depthwise separable convolution* and *inverted residual blocks*, offers high efficiency without sacrificing accuracy [3][4]. The *transfer learning approach* allows the model to leverage knowledge from large *datasets such as* ImageNet, then fine-tune it for smaller emotion datasets. This research focuses on developing a video-recording-based evaluation *pipeline* to test the performance of MobileNetV2 in classifying seven basic emotional expressions, to bridge the gap between static image models and video context.

2. RESEARCH METHODS

Study This was done with an experimental approach purpose evaluate the performance of the transfer learning- based MobileNetV2 model in recognizing facial expressions from image results, video extraction. The research process covers dataset selection, system pipeline design, implementation, use device soft supporters, as well as evaluation model performance through testing on test data.

2.1. Dataset

The main dataset used is *Karolinska Directed Emotional Faces* (KDEF), containing 4900 images of 70 subjects (35 men, 35 women) with seven basic emotions: *happy, sad, angry, fear, surprise, disgust,* and *neutral.* [5]This dataset is used to train and validate the model. To test, used video recording data of selected verbal interactions in a way selective from social media, then extracted become *frame* image to be more in accordance with the real condition.

2.2. Pipeline System

The pipeline study consists of several main :

1. Extraction *Frame*: Video recording has become a gathering *frame* image.
2. Detection: Every *frame* processed uses the algorithm *Hair Cascade* to extract the facial area.
3. *Preprocessing:* Detected faces changed their size to 224×224 RGB pixels and were normalized.
4. Classification Model: MobileNetV2 is used as the architecture base with *transfer learning* from *ImageNet*. Layers end modified into 7 classes expression. *Fine-tuning* was done to adjust the model to the KDEF *dataset*.
5. Evaluation : The model is evaluated using metric accuracy , *precision* , *recall* , F1-score, and *confusion matrix* .

2.3. Implementation

Experiment done using Python with the TensorFlow and Keras frameworks [6], as well as the OpenCV library for video processing. Training done on the Google Colab platform with standard hardware specifications (without an external GPU).

2.4. Data Evaluation and Analysis

Model evaluation is carried out with a measure of performance classification using metric accuracy, precision, recall, F1-score, and confusion matrix. In addition, the results predictions on analyzed video data continue for identifying pattern errors, for example, expressions that tend to be classified as neutral or influence factors external like lighting and facial angles. Analysis: This is used to understand the limitations of the model and provide a description of condition real moment system applied.

3. RESULTS AND DISCUSSION

3.1. Model Training Results

The training process showed consistent performance improvement from start to finish. Initial accuracy was still low, but continued to increase significantly after passing the 50th epoch. At the 200th epoch, the validation accuracy reached 0.8251, then increased to 0.8703 at the 240th epoch before the process was stopped at the 245th epoch.

The final results showed a training accuracy of 0.8673 and a validation accuracy of 0.8700 with low and stable loss values, indicating that the model had achieved optimal performance without symptoms of overfitting.

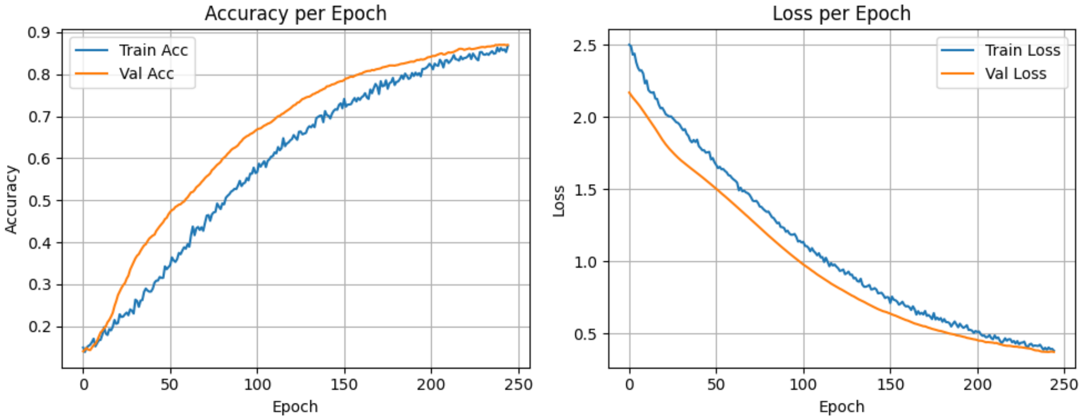

Figure 1Accuracy and Loss Curve

3.2. Model Evaluation Results

Model evaluation using the validation dataset produces a confusion matrix that shows the distribution of predictions on seven class emotions. From the results mentioned, it can be seen that expressions with strong visual characteristics, like happy to be recognized with okay, meanwhile, expressions subtle, such as anger and disgust, tend to be misclassified as neutral.

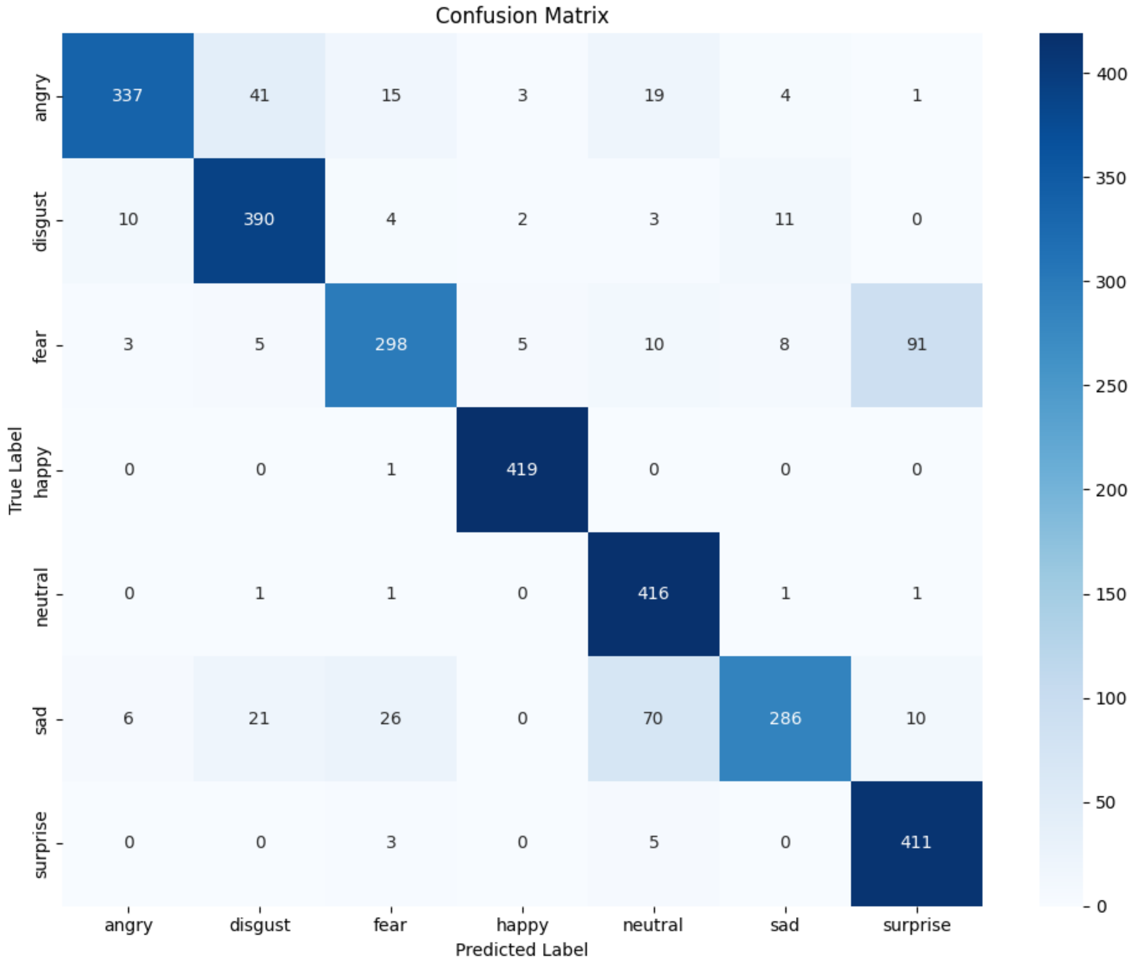

Figure 2Confusion Matrix

The following are the results of Other metrics of the confusion matrix that have been carried out :

a. Accuracy

$$Akurasi = \frac{609}{700} = 0.8703$$

b. Precision

$$Precision = \frac{240}{240 + 33} = 0.8784$$

c. *Recall*

$$Recall = \frac{240}{240 + 36} = 0.8703$$

d. F1- *Score*

$$F1 - Score = \frac{2 \times \ 0.8784 \ \times 0.8703}{0.8784 \ + 0.8703} = 0.8669$$

3.3. Test Results on Video Data

Test data in the form of video recordings of selected verbal interactions in a selective way, with 720p resolution and MP4 format. Each video is extracted into a frame, then taken sample of three frames per second is taken to reduce redundancy in the guard efficiency computing.

Table 1Number *Frame* Sampling

Video	Duration	Total Frame	Frame After Sampling
Angry	22 seconds	687	64
Disgust	15 seconds	461	43
Happy	10 seconds	306	28

All sampling frames are processed through the detection pipeline, face, preprocessing, and classification. Summary results prediction shown in Table 2.

Table 2Results of video test data classification

Video	Frame	Dominant Emotion	Percentage (%)	Average Confidence (%)	Distribution of Emotions
Angry	64	Neutral	75.0	51.71	Neutral (48), Angry (8), Disgust (8)
Disgust	43	Neutral	39.53	37.00	Neutral (17), Surprise (8), Angry (8), Disgust (6), Sad (4)
Happy	28	Happy	50.0	31.90	Happy (14), Disgust (7), Angry (4), Neutral (3)

This result shows the existence of model bias towards the neutral class, especially in the angry and disgust expressions. The average confidence in both videos was also low (<40%), indicating uncertainty predictions. On the other hand, in videos with happy expressions, more models are capable of recognizing expression-dominant, even though confidence remains relatively low. Findings. This confirms that the model is better at detecting expression with visual contrast features, while expression is still subtle and difficult to differentiate from a neutral face.

4. CONCLUSION

Study This successfully designing an evaluation pipeline video-based for classification expression using MobileNetV2 with a transfer learning approach. The pipeline includes stage frame extraction, face detection with Haar Cascade, image preprocessing, and classification expression.

Evaluation results show that the model achieves an accuracy validation of 87% with balanced performance on precision, recall, and F1-score. In testing a video-based system capable of identifying expression-dominant, although there is still a bias towards the neutral class, especially in expressions subtle (angry and disgust). On the other hand, the expression with more visual characteristics is clearer, like happier, easily recognized.

Thus , the goal study for linking static image models with video data has been achieved , even though there are still limitations that can be improved in further research.

5. SUGGESTION

As action further research furthermore can consider :

a. Use method detection more face-consistent, like MTCNN or MediaPipe, for extracting faces more stably.
b. Add variations in training data , in particular face Asian/Indonesian ethnicity, to reduce the bias of the KDEF dataset.
c. Apply temporal smoothing technique or aggregation between frames to reduce fluctuations prediction emotions in the video.
d. Explore other more advanced architecture light or more accurate, such as EfficientNet or Vision Transformer.
e. Test performance systems in various devices and environments for supporting HCI-based applications, device movement, or edge computing.

REFERENCE

[1] M. Chowdary, T. Nguyen, dan D. Hemanth, “Deep learning-based facial emotion recognition for human–computer interaction applications,” *Neural Comput Appl*, vol. 35, hlm. 23311–23328, 2021, doi: 10.1007/s00521-021-06012-8.

[2] W. A. Khan, A. U. Rahman, A. Zafar, A. Ashfaq, dan F. Karamat, “Facial Emotion Recognition: A Comparison of Classic and Novel Convolutional Neural Networks through Transfer Learning,” dalam *2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR)*, IEEE, 2024, hlm. 1–6.

[3] S. Kaur dan N. Kulkarni, “FERFM: An Enhanced Facial Emotion Recognition System Using Fine-tuned MobileNetV2 Architecture,” *IETE J Res*, vol. 70, hlm. 3723–3737, 2023, doi: 10.1080/03772063.2023.2202158.

[4] M. F. Nuryasin, C. Machbub, dan L. Yulianti, “Kombinasi Deteksi Objek, Pengenalan Wajah dan Perilaku Anomali menggunakan State Machine untuk Kamera Pengawas,” *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, &amp; Teknik Elektronika*, hlm., 2023, doi: 10.26760/elkomika.v11i1.86.

[5] E. Goeleven, R. De Raedt, L. Leyman, dan B. Verschuere, “The Karolinska Directed Emotional Faces: A validation study,” *Cogn Emot*, vol. 22, no. 6, hlm. 1094–1118, Sep 2008, doi: 10.1080/02699930701626582.

[6] N. Wiranda, H. S. Purba, dan R. A. Sukmawati, “Survei Penggunaan Tensorflow pada Machine Learning untuk Identifikasi Ikan Kawasan Lahan Basah,” *IJEIS (Indonesian Journal of Electronics and Instrumentation Systems)*, vol. 10, no. 2, hlm. 179, Okt 2020, doi: 10.22146/ijeis.58315.